DOCUMENT RESUME

ED 185 104 **TM 800 113**

AUTHOR Stonehill, Robert M.; Fishbein, Ronald L.

TITLE Summarizing the Results of Title I Evaluations--the Comparability of Achievement Gains.

PUB DATE Jun 79
NOTE 14p.; Paper presented at the Conference on Large-Scale Assessment (Denver, CO, June 11-14, 1979).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Achievement Gains; Achievement Tests; Comparative Analysis; Compensatory Education; Elementary Secondary Education; *Equated Scores; Federal Regulation; *Program Evaluation; Testing Problems; *Test Interpretation; Test Selection
IDENTIFIERS Aggregation (Data); Elementary Secondary Education Act Title I; *Normal Curve Equivalents; *Title I Evaluation and Reporting System

ABSTRACT

The Title I Evaluation and Reporting System (TIERS) was mandated to meet the requirements of the 1974 amendments to the Elementary and Secondary Education Act. It was hoped that TIERS would lead to a clear report of the effectiveness of Title I programs, based upon a synthesis of nationwide data. It has been extremely difficult to compare achievement data from all schools, local school districts, programs, states, and so on. It is not feasible for the federal government to mandate that one achievement test be used, since curricula are chosen at the local level. A certain amount of testing and calibrating must be performed before a set of tests are truly equated tests. The RMC Models' normal curve equivalent (NCE) scores have not reflected a common metric, as indicated by Jaeger's analysis of the Anchor Test Study. However, the NCE method shows promise as a metric which can be used for aggregation of Title I data from various sources. This method may be used to assign projects to rough categories of effectiveness (very effective to ineffective); if project variables are more controlled (such as statewide use of the same test), comparisons of project effectiveness would be even more valid. (GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED185104

SUMMARIZING THE RESULTS OF TITLE I EVALUATIONS — THE COMPARABILITY OF ACHIEVEMENT GAINS

Robert M. Stonehill Ronald L. Fishbein

U.S. Office of Education
Office of Evaluation and Dissemination

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Robert M. Stonehill

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM800 113

This paper was presented at the Conference on Large-Scale Assessment, Denver, CO, June 11 - 14, 1979.

SUMMARIZING THE RESULTS OF TITLE I EVALUATIONS —
THE COMPARABILITY OF ACHIEVEMENT GAINS

The passage of the Elementary and Secondary Education Act (ESEA) in 1965 was a landmark in that it was the first piece of Federal education legislation to provide substantial funds for categorical programs in compensatory education, as well as being the first example of major educational legislation which contained a provision for periodic evaluation. There was, in this case, a clear price for this kind of innovation — most of the painful lessons learned about evaluating the effects of Federal education legislation were learned at the expense of Title I.

Next year will mark the fifteenth anniversary since the passage of this seminal act, yet Congress has still to receive the first report, based on nationwide data, of the effectiveness of Title I programs in improving academic achievement. Hopefully, in that fifteenth year the U.S. Office of Education (USOE) will be reporting to the Congress on the effectiveness of the Title I program in increasing the reading, mathematics and language arts achievement of participants. Much of the basis for that report will be information obtained through the recently developed and implemented (and much debated) Title I Evaluation and Reporting System (TIERS).

TIERS was developed to meet the requirements of the 1974 amendments to ESEA. These amendments, drafted in response to the repeated failure of attempts to summarize meaningfully the results of state Title I evaluation reports, had several requirements, including the following:

- o the development of evaluation models which would yield data comparable on a statewide and nationwide basis,

- o the development of technical standards and procedures to be used when implementing these models,
- o the provision of technical assistance to assist State Education Agencies (SEAs) and local education agencies (LEAs) to implement the system.

The three basic evaluation strategies ("models") which constitute the TIERS have been widely disseminated, discussed and debated (e.g. Gamel, et al, 1975); hence, they will only tangentially be dealt with here. Suffice it to say that each model provides an estimate of the achievement growth attained by Title I participants above and beyond that attributable to only regular classroom instruction. The focus of this paper will be on the first of the above three requirements, the development of a system which yields comparable data.

Historically, very few items about Title I programs which were reported on in state Title I evaluation reports could be summarized across states (Gamel, et al, 1975; Thomas and Pelavin, 1976). Even the most basic kinds of data, such as the number of pupils who participated in Title I projects or the number of instructional personnel hired with Title I funds, were subject to such definitional differences across sites that the figures reported from place to place often had very different meanings. These definitional problems were corrected, for the most part, with the development and adoption of the TIERS, though even more uniformity of usage and meaning is still being sought. For example, the meaning of the word "project" still varies somewhat, though USOE is moving toward adopting a standard meaning for this term. Hopefully, USOE will be able to produce, with some confidence, nationwide estimates about the scope and range of Title I programs.

But the issues regarding the comparability of achievement impact information remain restive. In the past, the variance in test selection,

administration, analysis and reporting procedures encountered in state Title I evaluation reports completely precluded the possibility of summarizing this information in a way which would yield nationwide and statewide estimates of achievement growth due to the Title I program. During the initial stages of development of the TIERS, several strategies were contemplated to remedy that situation. In general, however, the more technically defensible the method for producing comparable data, the less practicable it would be.

One potential strategy is to select one instrument for national use. Achievement impact data produced at the LEA level would certainly be comparable across schools, districts, Title I programs, states, etc., if all local evaluators utilized the same test battery. Not only would it be a relatively straightforward procedure to aggregate impact data based on a common test, but all sorts of comparisons could be technically justified. The mean pretest or posttest status of different groups or grades could be compared; one could theoretically judge not only which programs made the greatest impact on the achievement of its students, but also which districts, or schools, or states, had the poorest (or highest) scoring students.

While there undoubtedly would be some interest in making comparisons of the sort described above, it is certainly not the intent of USOE, in developing and disseminating the TIERS, to encourage these sorts of comparisons. USOE's main purpose, as determined by Congress in Section 183 of ESEA, is to provide a system which would, while allowing for statewide and nationwide summaries of the effectiveness of the Title I program, also provide for valid evaluations at the local level. Mandating the use of a particular test would, in a sense, fulfill the first of these purposes,

but would almost certainly contravene the second. Prescribing a particular test would inevitably lead to prescribing a curriculum closely aligned with the features of that test, and the dictating of local curricula is not an activity appropriate for USOE. The strategy of mandating a particular achievement test is being used, in a limited manner, in the USOE-sponsored Sustaining Effects Study. In a sample of districts, student achievement is being measured with the Comprehensive Test of Basic Skills in a way that will allow for an assessment of their educational growth over a multi-year period. However, this is only a limited study, whereas the TIERS must apply to all Title I districts.

The operation of the Title I program relies on the LEA to conduct an assessment of the needs of its students, and to serve the most educationally disadvantaged students with programs designed to meet those needs. The valid evaluation of any such program would necessitate the selection of tests with content that approach, as far as possible, that of the respective Title I instructional program. In order to provide for this flexibility in test selection, compromises must be made with respect to the degree to which results produced through the application of different achievement tests will be comparable.

Given that the universal adoption of a single test is not desirable, the next best alternative, technically, would be to define a group of tests which have been equated. However, as Angoff (1971) says:

> "Equating, or the derivation of equivalent scores, concerns itself with the problem of unique conversions which may be derived only across test forms that are parallel, that is, forms that measure... the same psychological function."

This alternative then, that of providing a set of equated tests, is based fundamentally on the assumption that these tests measure the same construct, and in a similar manner. For a small set of standardized

achievement test batteries there is at least some evidence (e.g., the Anchor Test Study; Rentz and Bashaw, 1976) that the reading sections appropriate for grades 4 - 6 are equatable, either using equi-percentile methods or latent trait models.

Many points must be considered before reaching a decision as to the appropriateness of providing a set of fully equated tests, for grades 2 - 12. Whether the equating were to be performed using latent trait techniques, or classic equi-percentile methods, it is likely to be an expensive process. A particular test, or a test battery with a particular set of norms, must be deemed the anchor test, to which all others are then calibrated. To accomplish this, students would have to take pairs of tests (or at least sets of items from pairs of tests) simultaneously in order to equate the tests.

The positive features of such a system would include the fact that, to an extent, scores obtained from the different test forms included in such a system would be strictly comparable; that is, corresponding forms would yield equivalent percentile ranks on similar students. There are also some drawbacks. As initially structured, such a system would exclude a large number of tests -- many of which may be the tests of choice in particular localities. While such tests can be brought into the system, it is a non-trivial task, and it is based on the assumption that all such tests are psychometrically parallel (i.e. measure the abilities of students on a common laten trait). In practice, such will rarely be the case. Problems will arise whenever tests have not been explicitly constructed to be parallel with other forms; in fact, problems may occur whenever tests have different reliabilities even if parallel in content.

Another procedure for yielding comparable data was proposed by RMC Research Corp. when the TIERS was being developed. This procedure relies on the comparability of extant sets of norms. That is, under the assumption that the norms provided for any of the standardized achievement tests are fairly similar, then one only needs to adopt a common scale to yield comparable results. The NCE (normal curve equivalent) scale, a linear transformation of the z-score scale with a mean of 50 and a standard deviation of 21.06, has been adopted as such a common scale. To the extent that any test's percentile scores are truly rectangular, NCEs form a normalized, equal-interval scale. To the extent that the percentiles on Test A correspond to the percentiles that would be obtained should a similar group take Test B, then the NCE metric would be equivalent across those tests.

In fact, percentile ranks themselves are not very comparable for different tests. Consider the Anchor Test Study results, as re-analyzed and reported by Jaeger (1977). Differences between the Anchor Test Study norms and other publishers' norms vary across tests, with all exhibiting substantial variation between the NCEs derived, on the same group, for the different tests. This trend is most pronounced at the lowest ends of the score distributions, with five NCE differences between Anchor Test and publishers' norms common below the fifth percentile. Toward the middle of the score distributions, these differences are substantially smaller. Jaeger concludes as follows:

> " ... NCE conversion does not make the tests used in the Anchor Test Study equivalent, any more than would some other, more widely recognized, derived score. NCE scores do not reflect a common metric... It would appear that the Anchor Test Study norms and corresponding publishers' norms are somewhat different. They differ enough that this writer suggests that data on unequated achievement tests should not be pooled, regardless or NCE score conversion."

While it is clear that Jaeger's conclusion regarding the non-equivalence of scores obtained from the administration of different reading tests (actually, reading comprehension subtests) is valid, his conclusion regarding the pooling of information obtained from different tests is, in fact, not supported by his data. Table 1, which follows, is based on Table 9 in Jaeger:

Table 1
Mean Gains (NCE units) on 8 Reading Comprehension Subtests Corresponding to 5-Percentile Rank Gains on Anchor Test Study Norms for Grade 4 (from Jaeger,1977)

Anchor Test Study Percentile Rank Gain	Mean Gain in NCE Units on the Test Indicated							
	GMT*	ITBS	MAT	SAT	CAT	CTBS	SRA	STEP
1- 5	6.38	9.38	5.73	12.11	17.82	16.24	12.11	12.11
5-10	5.71	8.44	12.09	7.29	6.42	8.00	8.66	7.29
10-15	5.27	8.43	4.17	2.59	4.75	4.75	6.32	5.80
15-20	4.00	4.21	4.26	3.20	3.79	3.00	3.40	3.79
20-25	3.40	2.94	4.24	2.87	3.58	3.53	2.91	3.79
25-30	1.24	2.96	3.55	4.72	3.37	2.52	4.01	3.58
30-35	4.63	3.37	3.37	3.58	2.74	2.74	2.32	2.10
35-40	2.10	2.74	2.95	1.68	2.31	2.32	2.74	3.59
40-45	4.01	2.31	2.74	2.74	2.11	2.10	2.31	2.73
45-50	1.68	2.53	3.16	2.95	2.52	2.74	2.53	2.11
4.90	3.84	4.73	4.63	4.37	4.94	4.79	4.73	4.69

(Average NCE gains across percentile ranks)

Table 1 contains a comparison of the NCE gains, obtained using the publisher's norms for each of eight tests, that correspond to gains of five percentile points on the Anchor Test Study norms, at various levels of difficulty. Of course, five percentile point gains at different parts of the scale represent increasingly smaller increments in NCEs as one moves towards the 50tn percentile. In all but one case (a difference of 1.06 NCEs), the average NCE gain on a test, calculated across ability levels, is less than .53 NCEs different than the Anchor Test NCE average.

* The tests represented above are the reading comprehension subtests of, respectively, the Gates-MacGinitie, Iowa Test of Basic Skills, Metropolitan Achievement Test, Stanford Achievement Test, California Achievement Test, Comprehensive Test of Basic Skills, SRA Achievement Series, and Sequential Test of Educational Progress.

Table 2 contains the Anchor Test Study NCE gains corresponding to the intervals above, as well as the NCE gains that would be obtained by averaging across the 8 tests above; that is, the average aggregate NCE gain that would be obtained if the distribution of usage of the tests above was uniform.

Table 2

NCE Gains Obtained from Anchor Test Study Norms Compared to Average NCE Gains Based on Test Publishers' Norms for Grade 4 Reading Comprehension Subtests

Anchor Test Study Percentile Rank Gain	NCE GAINS		Discrepancy between Anchor Test Study norms and average of publishers' norms (absolute values)
	Anchor Test Study	Average of 8 Tests	
1- 5	14.4	11.49	2.91
5-10	7.6	7.99	.39
10-15	5.2	5.26	.06
15-20	4.1	3.71	.39
20-25	3.5	3.41	.09
25-30	3.2	3.24	.04
30-35	2.9	3.11	.21
35-40	2.8	2.55	.25
40-45	2.7	2.63	.07
45-50	2.6	2.53	.07
Average NCE gains (assuming uniform occurrence)	4.90	4.59	

Given that the pattern of usage is fairly uniform for all of the eight tests included in the Anchor Test Study, it is evident that NCE gains, when aggregated across different ability levels and across different tests, will approximate closely the NCE gains that would have been obtained had only the Anchor Test Study norms been utilized. In fact, since only a small proportion of Title I projects will have NCE gains based on percentile changes occurring in the lowest five percentile ranks (where the largest discrepancy by far is found), the estimates obtained by pooling results should be even more accurate than implied in Table 2.

One further examination of the differences between NCE aggregates and Anchor Test Study results is presented in Table 3, which displays weighted NCE gain aggregates. Table 3 represents the combining of data from projects all showing 5 percentile point gains on the Anchor Test; however, rather than being uniformly distributed across achievement levels, they are normally distributed around projects which have moved students from the 20th to the 25th percentile.

Table 3

Weighted NCE Gains from the Anchor Test Study Norms Compared to Average NCE Gains Based on Test Publishers' Norms for Grade 4 Reading Comprehension Subtests

Anchor Test Study Percentile Rank Gain	Weighting of project gains to simulate normal distribution of ability levels	NCE GAINS	
		Anchor Test	Average of 8 Tests
1- 5	.0179	14.4	11.49
5-10	.0489	7.6	7.99
10-15	.1173	5.2	5.26
15-20	.1980	4.1	3.71
20-25	.2358	3.5	3.41
25-30	.1980	3.2	3.24
30-35	.1173	2.9	3.11
35-40	.0489	2.8	2.55
40-45	.0144	2.7	2.63
45-50	.0035	2.6	2.53
	Weighted NCE gains =	4.04	3.93

If the results from Title I project evaluations which manifested 5 percentile point gains (based on Anchor Test norms) were aggregated, the NCE gains would differ from those which would be obtained only by using Anchor Test norms by about .4 NCEs if projects were uniformly distributed across percentile ranks, and would differ by only about .1 NCEs if they were normally distributed around projects whose means were in the 20th to the 25th percentile (typical for Title I projects).

It seems reasonable to assume that this pattern of convergence would be obtained for other grades, and for subtests other than reading achievement

(particularly, for Total Reading).

Though NCEs appear to hold some promise as a metric which can be used for aggregation of data from various sources, there is certainly equally compelling evidence that they should not be interpreted as a measure which allows for strict comparisons of project evaluation outcomes. At best they may be useful for assignment of projects to some rough categories of effectiveness (very effective to ineffective). However, as some variables are controlled (for instance, if all projects in an area used the same test), the system may allow for more valid comparisons of project effectiveness.

The introduction of the NCE metric for reporting and aggregation purposes has led to some interesting re-examinations of test results. Where it had previously been extremely uncommon to question the meaning or validity of scores which had a familiar mantle (particularly grade equivalents), the adoption of NCEs was accompanied with close attention to the meaning of the new metric, and to the issues of comparability which TIERS attempted to address. In fact, NCEs are no better or worse than any other standard score scale, and many of the criticisms leveled against NCEs would more appropriately have been directed toward the test publishers' norms themselves, which are often filled with imprecision and inconsistency, especially at the very low end of the distributions of achievement. With or without conversion to the NCE metric, such entities as expanded standard score scales (which link levels of a test to a common scale) and very low or high percentile ranks do not behave as impeccably as we would like.

A compelling reason for choosing a metric like the NCE, which is derivable from information inherent in the norms of any particular

normed instrument and not relying on any external equating efforts, is the freedom to choose the test most appropriate for measuring the goals of a local Title I project. The evaluator is not limited to a small set of tests which has previously been equated; so long as norms exist, or are made available, a test can fulfill the requirements of the Title I evaluation system. However, the evaluator must implicitly believe that the norms obtained for the test of choice are valid, and representative of the type of children who participate in Title I within the LEA, or SEA, selecting that test. In a system based on a set of equated, or calibrated instruments, this burden of responsibility shifts from the individual test publisher to the person who has done the actual equating.

In conclusion, a single test or a set of tests which have been calibrated to a single set of norms, will yield information comparable in two ways -- actual achievement levels as well as achievement gains can be examined. However, a system based on equated tests can only give the appearance or providing comparable data; in fact, if the tests entered into that system are not very similar types of instruments, results will be inconsistent and potentially misleading.

In adopting a system based on the NCE as an aggregation metric, USOE believes it has reached a defensible solution to the problem of comparing and pooling data on achievement gains, given that compromises had to be made between technical and practical constraints. However, difficult issues pertaining to data aggregation, both technical and philosophical, still remain.

To quote Angoff (1971) once again:

> "Statistical solutions are fundamentally no more precise than the data they are based on and no more defensible than the methods used to derive them."

REFERENCES

Angoff, W.H. "Scales, Norms, and Equivalent Scores." In Educational Measurement, Second Edition. R.L. Thorndike, ed. Washington, D.C.: American Council on Education, 1971.

Gamel, N.N., Tallmadge, G.K., Wood, C.T., & Binkley, J.L. State ESEA Title I Reports: Review and analysis of past reports, and development of a model reporting system and format. Mountain View, CA: RMC Research Corp., October, 1975.

Jaeger, R.M. On Combining Achievement Test Data Through NCE Scaled Scores. Research Triangle Park, NC: Research Triangle Institute, 1977.

Lord, F.M. and Novick, M.R. Statistical Theories of Mental Test Scores. Reading, MA: Addison-Wesley, 1968.

Rentz, R.R. and Bashaw, W.L. Equating Reading Tests with the Rasch Model. Athens, GA: University of Georgia, 1975.

Thomas, T.C. and Pelavin, S.H. Patterns in ESEA Title I reading achievement. Menlo Park, CA: Stanford Research Institute, March, 1976.